ESTIMATING PROBABILITIES*

WILLIAM M. BRELSFORD AND RICHARD H. JONES
Department of Statistics, The Johns Hopkins University, Baltimore, Md.

ABSTRACT

Probability prediction is defined as the estimate of the conditional probability distribution of the future given the past and present. Of principal interest are probability predictions for dichotomous random variables, i.e., variables assuming the value 1 (success) or 0 (failure), in which case the prediction is simply an estimate of the probability of success.

Given certain assumptions on the distribution of the independent variables, the probability of a success is shown to be a logit curve, a type of sigmoid. The maximum likelihood estimators of the parameters of the logit are obtained and compared with estimators obtained by non-linear regression techniques—both classical and a newer recursive method. A comparison is also made with a linear regression line, showing the conditions under which it provides an acceptable predictor. The comparisons are made by analysis of several types of simulated data, as well as by analysis of meteorological data.

Binary events with continuous underlying variables (e.g., daily temperature exceeding its mean by k degrees) are also considered. The probability prediction is made by estimating the conditional distribution of the continuous variable and integrating over the region of interest.

1. INTRODUCTION

Let z be a random variable and x an observed value of an associated vector variable. We are interested in the probability structure of z. All information regarding z given by the observation is contained in the conditional probability distribution, $F(z|x)$. We therefore define a probability estimate to be an estimate $\hat{F}(z|x)$ of the conditional probability distribution. In particular, this paper is concerned with probability estimation for dichotomous random variables, i.e., $z=1$ (success) or 0 (failure). For example, z might represent the occurrence or non-occurrence of rainfall.

2. THE LOGIT MODEL

Let x be a random variable chosen from one of two populations, f_0 and f_1, leading to $z=0$ and $z=1$, respectively, and let $p=\Pr(z=1)$ and $q=1-p$. If $f_i=N(\mu_i, \Sigma)$, it may be shown that

$$\Pr(z=1|x)= \left\{1+\frac{q}{p}\exp-\left[x-\frac{1}{2}(\mu_1+\mu_0)\right]'\Sigma^{-1}[\mu_1-\mu_0]\right\}^{-1}, \quad (2.1)$$

which is of the form

$$E(z|x)=\frac{1}{1+e^{-(\alpha+x'\beta)}}. \quad (2.2)$$

This is the logistic response function, or logit, a symmetric sigmoid curve. It appears to be a reasonable model since, as a smooth function of x, $\Pr(z=1|x)$ is bounded between 0 and 1 and approaches these values as limits as $x\to\pm\infty$.

3. THE DISCRIMINANT FUNCTION LOGIT ESTIMATOR

This formulation of the problem was discussed by Cornfield, Gordon, and Smith [1], who use maximum likelihood estimates of α and β. Thus,

$$\begin{aligned} b&=S^{-1}(\bar{x}_1-\bar{x}_0) \\ a&=-\log(n_0/n_1)-\tfrac{1}{2}(\bar{x}_1+\bar{x}_0)'b. \end{aligned} \quad (3.1)$$

S is the pooled estimate of the common covariance matrix Σ. The exponent in (2.1) is seen to be the linear discriminant function and the estimates (3.1) are based on the sample discriminant function, so we will call this estimator the discriminant function logit estimator (DFLE).

4. OTHER LOGIT MODELS

Without the assumption of equal variances in the two cases, the exponent in (2.1) becomes the quadratic discriminant function, and its sample version is the difference between the Mahalanobis squared distances from the observation to each sample. This yields a quadratic logit model.

Suppose x has univariate gamma distributions, with parameters dependent on $z=0$ or 1. Then the model is of the form

$$E(z|x)=\frac{1}{1+e^{-(\alpha+\beta x+\gamma\ln x)}}. \quad (4.1)$$

*This research was based on part of the first author's doctoral dissertation submitted to the Johns Hopkins University, and was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, U.S. Air Force, under AFOSR Contract No. AF49(638)-1302.

Thus, for gamma variables, the model includes both the variable x and its logarithm.

5. REGRESSION MODELS

Often the distributions of x given $z=0$ and $z=1$ are not known, or x is not a random variable at all. In such cases $\Pr(z=1|x)$ may be estimated by regression methods.

The simplest regression model would be

$$E(z|x)=x'\beta, \tag{5.1}$$

where $x_1 \equiv 1$. Here the estimates $\hat{z}=x'b$ will not be constrained to lie between 0 and 1. This is similar to the Regression Estimation of Event Probabilities (REEP) procedure of Miller [6]. A second model is the logit discussed above, i.e.,

$$E(z|x)=\frac{1}{1+e^{-x'\beta}}. \tag{5.2}$$

These models will be compared in section 6.

The logit parameters may be estimated by non-linear least squares. The function (5.2) is expanded in a Taylor series about an initial estimate b, with only the linear term retained. This leads to the "linearized" normal equations

$$X'[\hat{P}\hat{Q}]^2X\widehat{(\beta-b)}=X'\hat{P}\hat{Q}(z-\hat{z}), \tag{5.3}$$

where

$$\hat{z}=\frac{1}{1+e^{-x'b}},$$

$$\hat{P}=\begin{bmatrix} \ddots & & 0 \\ & \hat{z}_i & \\ 0 & & \ddots \end{bmatrix},$$

and

$$\hat{Q}=I-\hat{P}.$$

The parameters of logistic distributions can also be estimated using maximum likelihood by obtaining "linearized" normal equations

$$X'\hat{P}\hat{Q}X\widehat{(\beta-b)}=X'(z-\hat{z}), \tag{5.4}$$

which appears to be a weighted version of (5.3), except that the weights $(\hat{P}\hat{Q})^{-1}$ are functions of the estimate b. We will use this form of the normal equations. β is then estimated by iteration, i.e., by using $\widehat{(\beta-b)}$ obtained from (5.4) to improve the estimate b, and iterating until the correction term becomes sufficiently small. Convergence is guaranteed only if the initial estimate b is "close enough" to the true value β.

Walker and Duncan [8] use a recursive technique to estimate β. Given an estimate b_{n-1} based on $n-1$ observations, the estimated covariance matrix $\Sigma_{n-1}=\widehat{\mathrm{Var}}\ (b_{n-1})$, and the n-th observation (x_n, z_n), the estimates of β and Σ may be updated as follows:

$$\Sigma_n=\Sigma_{n-1}-\frac{\Sigma_{n-1}x_nx_n'\Sigma_{n-1}}{x_n'\Sigma_{n-1}x_n+[\hat{z}_n(1-\hat{z}_n)]^{-1}}$$

$$b_n=b_{n-1}+\frac{\Sigma_{n-1}x_n(z_n-\hat{z}_n)}{x_n\Sigma_{n-1}x_n+[\hat{z}_n(1-\hat{z}_n)]^{-1}}, \tag{5.5}$$

where

$$\hat{z}_n=\frac{1}{1+e^{-x_n'b_{n-1}}}.$$

6. COMPARISON OF LINEAR AND LOGIT MODELS

Analysis of simulated data has been carried out to investigate the performance of the linear model (5.1) and the logit model (5.2) under varying conditions. Random variables x and y were drawn from a standard bivariate normal distribution with correlation ρ. The dichotomous variable z was then defined as

$$z=1 \text{ if } y>c$$
$$=0 \text{ if } y<c.$$

Thus the true form of the response curve was the cumulative normal (probit).

Both the linear and logit curves were estimated by the recursive technique mentioned previously. A sample size of 1,000 was used to assure convergence.

Figure 1 shows the results of a typical run. The correlation between x and y in this case was 0.9. The cut value c was 1, giving an unconditional probability of a success of 0.1588.

Figure 2 demonstrates the effects of the correlation and the cut value on the fitted curves. The logit provides a better approximation, especially in the tails, as c departs from zero. The straight line, however, becomes a reasonably close approximation over a range as $|\rho|$ decreases.

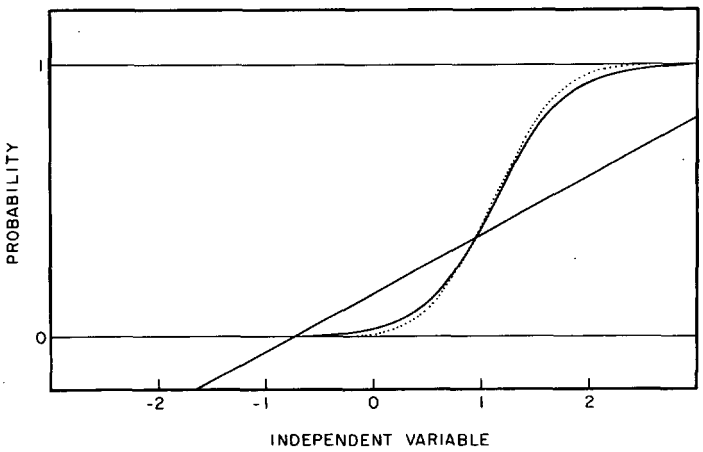

FIGURE 1.—Comparison of linear and logit probability estimators for $\rho=0.9$, $c=1$, $n=1{,}000$. The true probability function (dotted line) is a normal c.d.f.

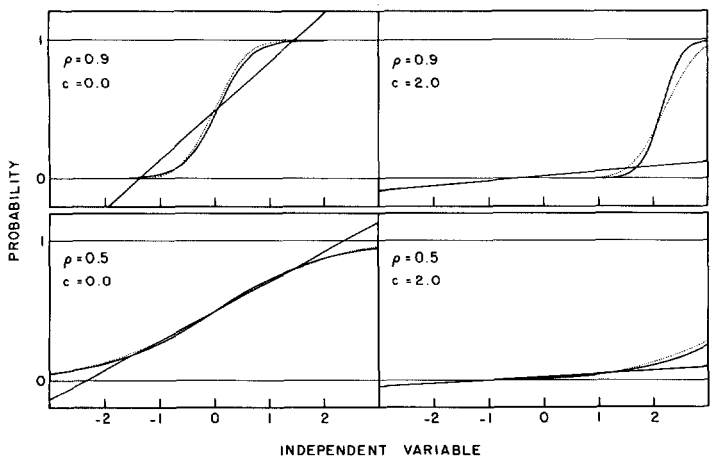

FIGURE 2.—Linear and logit probability estimators for two values of ρ and c with $n=1{,}000$. The true probability function (dotted line) is a normal c.d.f.

7. COMPARISON OF LOGIT ESTIMATION METHODS

A second set of simulations was performed to compare the DFLE, converged iterative, and one-pass recursive logit estimators. A sample size of 100 was used, and the converged iterative estimate was considered to be the standard of comparison, since it provides the best fit to the data.

The results of two runs made on data satisfying the DFLE assumptions are shown in figure 3. Here

$$p=Pr(z=1)=0.8,$$

$$f_0(x)=N(0,\ 1),$$

and

$$f_1(x)=N(1,\ 1).$$

The DFLE is seen to be very close to the iterative solution in both cases.

Figure 4 shows curves resulting from analysis of gamma and uniform data, with the true response curve being the logit in each case.

In order to illustrate differences in rate of convergence for the iterative and recursive methods, consider again the data which resulted in the upper graph in figure 3. As our starting value in each case we take a horizontal line $E(z|x)=\hat{p}$. The proportion of successes in this case was 0.83. We are interested in the value of the coefficients after each iteration. In the recursive case the regression coefficients which resulted from the previous pass through the data were used as initial estimates for the next pass. For each pass the covariance matrix of the regression coefficients was reset to the initial value used in the first pass. The results shown in table 1 show that the recursive method provides a reasonable estimate in one pass, and that the iterative method is particularly sensitive to

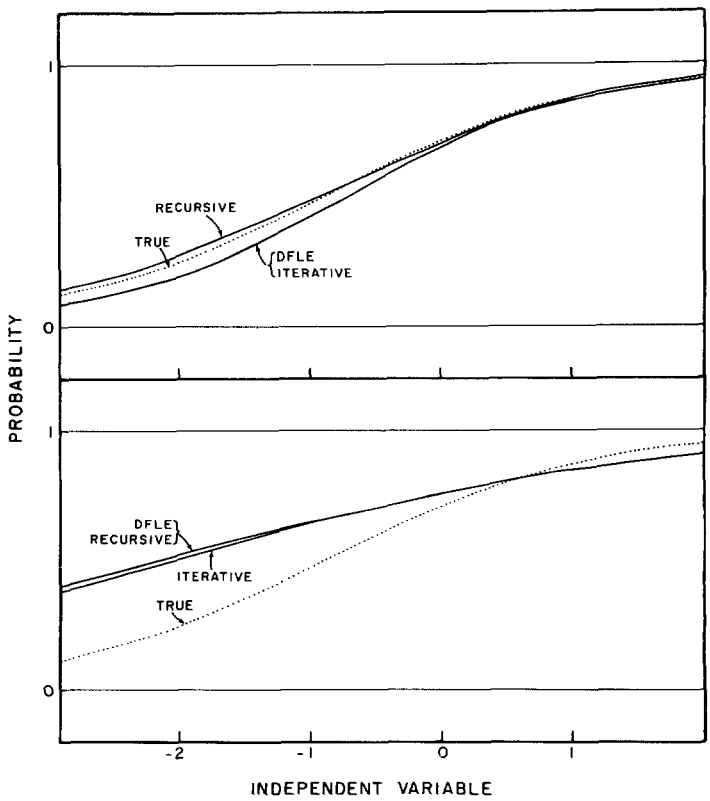

FIGURE 3.—Probability estimators for two sets of data satisfying the DFLE conditions, with $n=100$.

TABLE 1.—*Convergence rates for iterative and recursive logit estimators for various starting values,* $\hat{p}$

$\hat{p}$	Iteration	Iterative		Recursive	
		b_0	b_1	b_0	b_1
0.40	0	−0.405	0.000	−0.405	0.000
	1	0.791	0.582	0.803	0.902
	2	0.786	0.953	0.785	1.036
	3	0.785	1.073	0.779	1.048
0.76	0	1.153	0.000	1.153	0.000
	1	0.753	0.765	0.845	0.912
	2	0.784	1.022	0.786	1.037
	3	0.785	1.081	0.779	1.048
0.83	0	1.586	0.000	1.586	0.000
	1	0.573	0.989	0.853	0.918
	2	0.775	1.061	0.785	1.037
	3	0.785	1.083	0.779	1.048
	4	0.785	1.084	0.778	1.049
0.90	0	2.197	0.000	2.197	0.000
	1	−0.167	1.551	0.863	0.915
	2	0.853	0.956	0.786	1.037
	3	0.786	1.078	0.779	1.048
0.95	0	2.944	0.000	2.944	0.000
	1	−2.589	2.939	0.875	0.893
	2	5.325	−3.172	0.787	1.034
	3	−17.316	12.668	0.779	1.048
0.99	0	4.595	0.000	4.595	0.000
	1	−25.993	14.103	0.903	0.831
	2	4025.930	−2208.760	0.791	1.029
	3	--------	----------	0.779	1.047

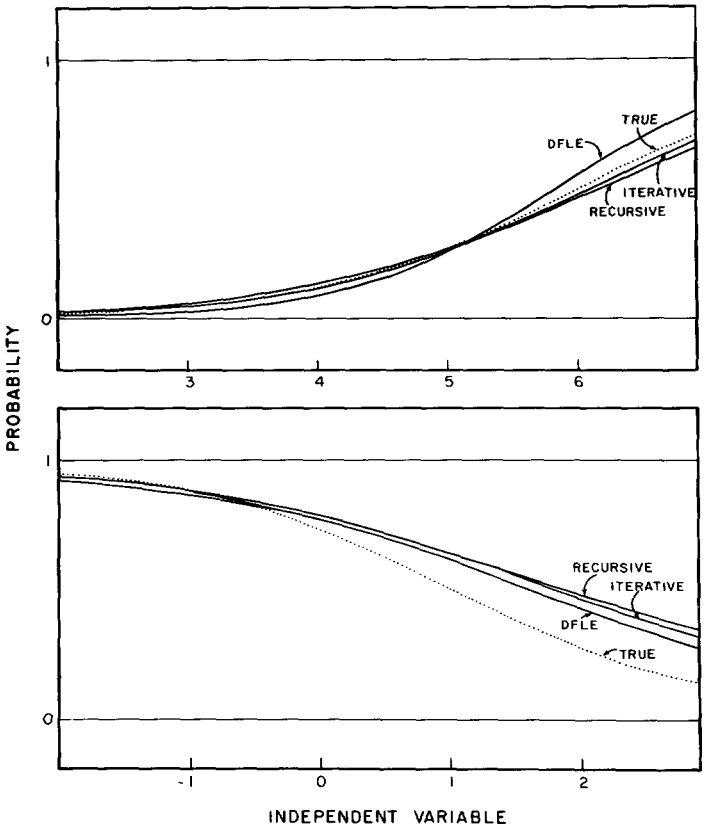

FIGURE 4.—Probability estimators, where the independent variable has a gamma distribution (top) and uniform distribution (bottom), with $n=1{,}000$. The true probability function is a logit.

starting values. A starting value too near 1 causes divergence.

In choosing an estimation method to use for a particular problem, several factors should be considered. The DFLE has the advantage of being the easiest to calculate and provides good estimates if the model is correct. The iterative method, if it converges, converges to the maximum likelihood estimate. Recursive estimation provides an updating feature which is especially useful in real-time situations. This allows the data to be used as "independent data" to judge performance in addition to its role in determining the coefficients. The recursive estimate converges reasonably well in one pass for a moderate sample size so that the data need not be retained. This is important in computer applications where data storage is limited or expensive. Finally, this estimate is relatively independent of the intial values of the coefficients.

Combinations of the three methods should be considered. The DFLE gives a good starting point, if some data are available. Iterative regression may be used to improve the initial estimate if the DFLE assumptions are not satisfied. Recursive estimation might then be used to update as more observations are obtained. At some point one may wish to stop and iterate over the data up to that point, using either iterative or recursive methods.

8. UNDERLYING VARIABLES

Let z be a dichotomous random variable with an observable underlying variable y, i.e.

$$z=1 \text{ if } y\epsilon R$$
$$=0 \text{ otherwise,}$$

where R is some known region in y-space. For example y might be the ceiling height and R the set of all heights below 600 ft. Given x, a probability estimate for z can be made based on the estimated conditional distribution of y given x, since

$$E(z|x)=\int_R dF(y|x). \tag{8.1}$$

In meteorological problems these underlying variables will sometimes be approximately normally distributed. Others, such as amount of rainfall, can be transformed so as to be approximately normal. If we fit the linear model

$$y=X\beta+\epsilon, \tag{8.2}$$

then, given x, the residual $y-x'b$ will be normally distributed with mean zero and variance $(1+x'(X'X)^{-1}x)\sigma^2$. The problem is now reduced to that of estimating the normal c.d.f. at one or more points. For moderate or large sample sizes σ^2 may be replaced by its estimate s^2 and the integration may then be performed.

9. THE k-CHOTOMOUS LOGIT MODEL

Suppose each trial results in the occurrence of one of k mutually exclusive events. A general extension of the logit to this problem was given by Cox [2]. Let

$$z=(z_1, \ldots, z_k)'$$

be a vector denoting the result of the trial, with $z_m=1$ and $z_j=0$ for $j\neq m$. Let

$$E(z_i|x)=\frac{e^{-x'\beta_i}}{\sum_{j=1}^{k} e^{-x'\beta_j}}. \tag{9.1}$$

Since $\sum_{j=1}^{k} z_j=1$ we may arbitrarily set $\beta_1=\underset{\sim}{0}$. For $k=2$ this reduces to the dichotomous model (5.2). Estimation may be carried out by methods generalized from the dichotomous case.

If the events have a natural ordering, it is possible to reduce the k-chotomous problem to $k-1$ dichotomous problems by grouping the events. In this case only the slope coefficient is allowed to vary with the different groupings (see Walker and Duncan [8]).

10. SCORING PROBABILITY FORECASTS

Probability estimates or forecasts for the k-chotomous model may be compared by a loss function, $h(z, \hat{z})$. The Brier Score used in meteorology is essentially mean square error, i.e., it assigns a loss of $\sum_{j=1}^{k}(z_{ij}-\hat{z}_{ij})^2$ to the i-th trial (the subscript i will be used to denote the trial and j the event). The average loss for this loss function is minimized by the non-linear least squares estimate obtained from (5.3).

Another natural loss function derived from information theory assigns a loss of $-\log \hat{z}_{im}$, where m is the event which occurred. Loss functions of this type have been used by Holloway and Woodbury [5], Suzuki [7], and others. It is important to note that for parametric probability models, this loss function is minimized by the maximum likelihood estimate of the parameters. The log likelihood of the sample is

$$L=\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log p_j(x_i), \tag{10.1}$$

where

$$p_j(x_i)=E(z_{ij}|x_i). \tag{10.2}$$

Maximum likelihood chooses the estimate of β which maximizes (10.1), thereby minimizing

$$-\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log \hat{z}_{ij}.$$

A loss function for which the average loss for the sample is minimized by the maximum likelihood estimator is thus

$$h(z_i,\hat{z}_i)=-\sum_{j=1}^{k} z_{ij}\log \hat{z}_{ij}=-\log \hat{z}_{im}. \tag{10.3}$$

For objective probability estimates, we prefer the information loss function for the following reasons: (1) For a finite sample, the average loss is minimized by maximum likelihood estimation. (2) The expected loss is minimized by the true probabilities and is equal to the entropy of the distribution. (3) Estimates are constrained to the range $0\leq\hat{z}\leq 1$. A probability prediction of zero is unacceptable if the event occurs, since the loss would be $+\infty$.

11. METEOROLOGICAL EXAMPLES

Several types of meteorological data have been analyzed to further evaluate the practical effectiveness of the logit and underlying variable probability predictors. These predictors are compared with those obtained by linear regression and persistence. The persistence predictor, which is often used as a "minimum" standard of comparison in probability forecasting, is simply the outcome at the preceding time point,

$$\hat{z}_t=z_{t-1}. \tag{11.1}$$

Since meteorological events often occur in runs, persistence frequently appears to perform reasonably well; clearly, to be worthwhile, a proposed predictor must do better.

The loss function used to compare predictors is mean square error,

$$h(z,\hat{z})=\frac{1}{n}\sum_{t=1}^{n}(z_t-\hat{z}_t)^2, \tag{11.2}$$

since the information loss function of section 10 cannot be used to score persistence. If persistence were replaced by the zero-information probability predictor, $\hat{z}=\bar{p}$, where $\bar{p}$ is an estimate of the unconditional probability vector, as the standard of comparison, the information loss function could be used.

In the following examples regression coefficients are estimated by the recursive techniques discussed earlier, enabling the sample to be used as an "independent sample" to measure performance. At each time point a prediction is made based on past data. The observation obtained at that point is compared with the prediction, and then used to update the estimates of the coefficients. In order to allow the coefficients to stabilize, the first 100 data points were excluded from the mean square error calculations.

In the first example, hourly weather data from the Atlantic City, N.J., airport were used to predict the probability of precipitation. Atmospheric pressure and dew point depression were the only relevant predictor variables available; the squares and cross product of these variables were also used. The first analysis, conditional on precipitation the preceding hour, showed the logit to be slightly better than linear regression and both better than persistence. The second, conditional on no rainfall, resulted in approximately equal results for the three predictors. In neither case were any of the regressors statistically significant. Deleting these regressors results in a Markov chain model, which is commonly used for predicting precipitation probabilities (see, for example, Eriksson [3], Feyerherm and Bark [4]).

Attention was shifted to temperature data for Central Park in New York City. This consisted of daily minimum and maximum temperature readings for a 30-yr. period, 1931–60. We wished to predict the probability that the maximum would exceed its daily mean or the minimum fall short of its daily mean by k degrees. Estimates of the daily means were obtained by averaging over the 30 yr. A bivariate series $\{x_t\}$ was obtained by subtracting out these means.

The specific problem considered in this example was of predicting the probability that the minimum temperature $x_{2,t}$ would be 5° or more below its daily average, i.e.,

$$z_t=1 \text{ if } x_{2,t}<-5 \tag{11.3}$$

$$=0 \text{ otherwise.}$$

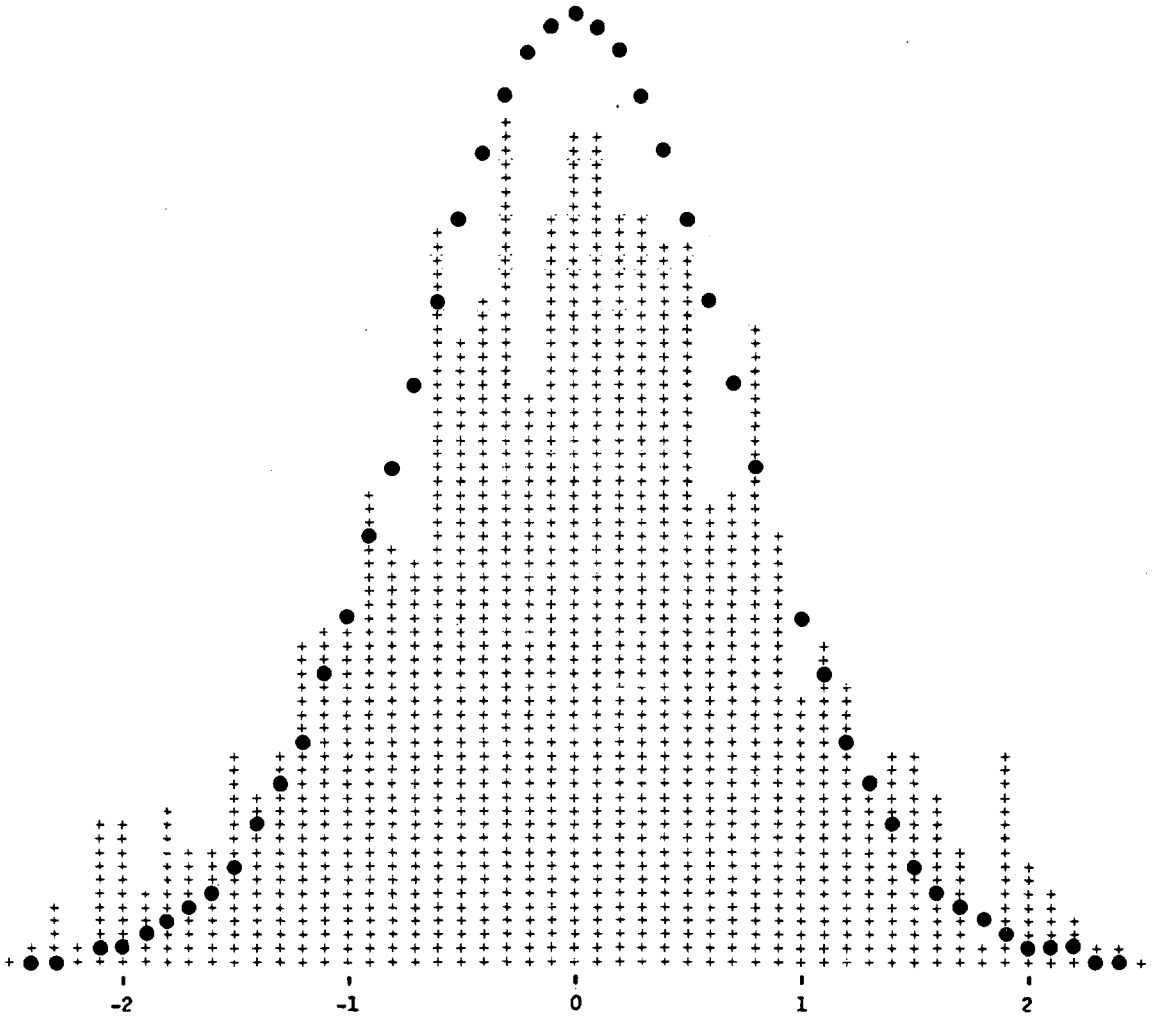

FIGURE 5.—Distribution of residuals for underlying variable probability predictor, normalized by current parameter estimates, with superimposed standard normal curve.

TABLE 2.—*Results of experiment on minimum temperature*

	$z_{t-1}=0$		$z_{t-1}=1$	
Sample size	800		600	
M. S. E.				
Logit	0.1024		0.2018	
Linear	.1042		.2041	
Persistence	.1243		.4200	
Variable	Coeff.	t	Coeff.	t
x_0	−1.561	—	−1.403	—
$x_{1,t-1}$	−0.216	−5.72	−0.242	−6.56
$x_{2,t-1}$	−0.007	−0.32	−0.095	−4.96
$x_{1,t-2}$	0.032	1.01	0.187	6.14
$x_{2,t-2}$	0.020	0.92	0.005	0.25
$x_{1,t-3}$	0.002	0.06	−0.047	−1.90
$x_{2,t-3}$	−0.014	−0.65	−0.009	−0.51

The variables used as regressors were the maximum and minimum temperatures for the preceding three days. Two conditional logit regression analyses were performed with the results shown in table 2. This and similar studies (e.g., predicting the probability that the maximum temperature exceeds its average by 10°) show the logit to be consistently better than the linear model.

Since the time series in question is approximately Gaussian, the underlying variable probability predictor should perform well in this case. This was done, using three lags of the bivariate series as predictor variables. The prediction was made at each point using current estimates of the parameters of the distribution obtained by recursive least squares. The distribution of the residuals, normalized by the current parameter estimates, is shown in figure 5; the first 100 points of the 1000 used are omitted. The underlying variable probability predictor had a mean square error of 0.1291, compared to 0.1956 for persistence.

REFERENCES

1. J. Cornfield, T. Gordon, and W. W. Smith, "Quantal Response Curves for Experimentally Uncontrolled Variables," *Bulletin*, International Statistical Institute, vol. 38, part III, 1961, pp. 97–115.
2. D. R. Cox, "Some Procedures Connected with the Logistic Qualitative Response Curve," *Research Papers in Statistics* (F. N. David, Ed.), John Wiley and Sons, London, 1966.
3. B. Eriksson, "A Climatological Study of Persistency and Probability of Precipitation in Sweden," *Tellus*, vol. 17, No. 4, Nov. 1965, pp. 484–497.
4. A. M. Feyerherm and L. D. Bark, "Statistical Methods for Persistent Precipitation Patterns," *Journal of Applied Meteorology*, vol. 4, No. 3, June 1965, pp. 320–328.
5. J. L. Holloway, Jr. and M. A. Woodbury, "Application of Information Theory and Discriminant Function Analysis to Weather Forecasting and Forecast Verification," Technical Report No. 1, Meteorological Statistics Project, University of Pennsylvania, 1955.
6. R. G. Miller, "Regression Estimation of Event Probabilities," Technical Report, Travelers Research Center, 1964.
7. E. Suzuki, "Weather Forecast and Entropy in Information Theory," *Papers in Meteorology and Geophysics*, Tokyo, vol. 9, No. 2, Dec. 1958, pp. 51–62.
8. S. H. Walker and D. B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, vol. 54, 1967, pp. 315–327.

[*Received April 19, 1967; revised May 25, 1967*]